New Mexico Data Comparison

Joe Champion, Thea Sukianto

February 23, 2020

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
```

Prepare Data

We append flags to the `candidate` data with the `clean_location_data` function from the `animaltracker` package.

If the `Rate` is greater than 84, we append a `RateFlag`.

If the `Course` is greater than or equal to 100, we append a `CourseFlag`.

If the `Distance` is greater than or equal to 840, we append a `DistanceFlag`.

In the cleaning process, observations with a `DistanceFlag`, or 2+ flags are removed.

However, the data is left unchanged in this case for comparison purposes.

```
candidate <- read.csv("df_candidate.csv", stringsAsFactors = FALSE)
correct <- read.csv("df_correct.csv", stringsAsFactors = FALSE)
print(nrow(candidate))
```

```
## [1] 167901
```

```
print(nrow(correct))
```

```
## [1] 167901
```

We use the `dplyr` package to join the `candidate` and `correct` data on `Cow` and `Index`.

```
join <- dplyr::full_join(candidate, correct, by=c("Cow", "Index"))
print(nrow(join))
```

```
## [1] 168413
```

There are approximately 500 more observations in the joined data than there are in each individual dataset.

Analysis

First, we determine which observations in `candidate` are to be kept according to the `clean_location_data` function.

```
join <- join %>%
  dplyr::rename(Keep.y = Keep,
                DistFlag.x = DistanceFlag,
                DistFlag.y = DistFlag) %>%
  dplyr::mutate(Keep.x = 1*(TotalFlags.x < 2 & !DistFlag.x))
```

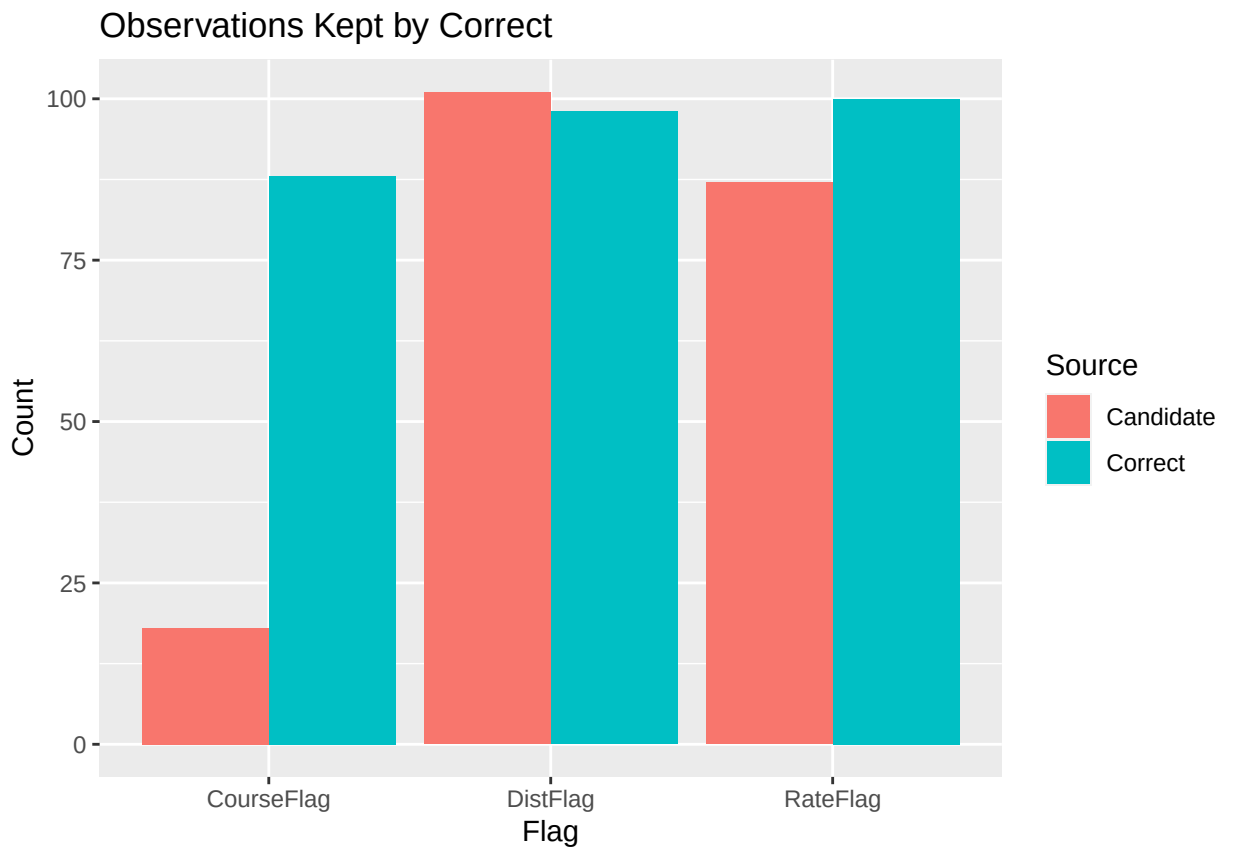
There are 165039 observations that both `correct` and `candidate` keep and 2935 that both discard.

However, `correct` discards 337 that `candidate` would not and `candidate` discards 101 that `correct` would not.

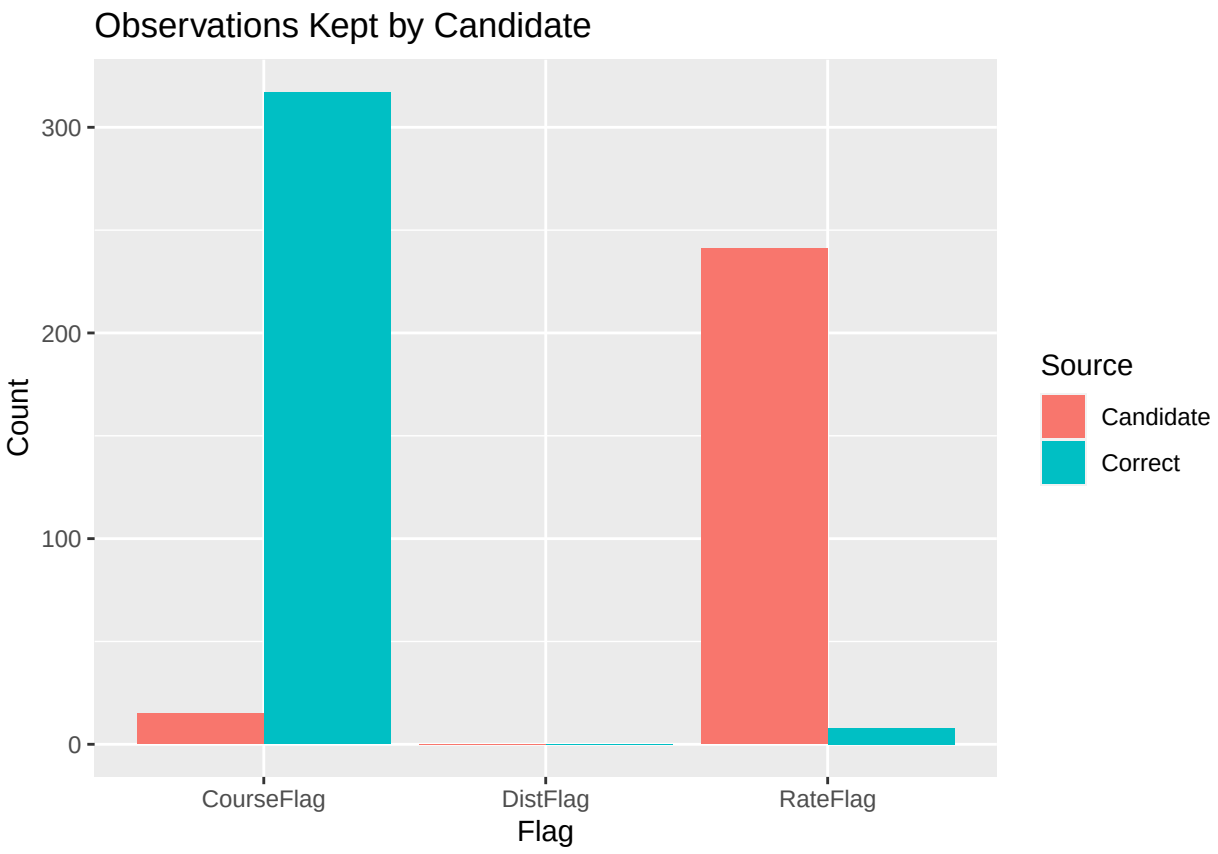
```
join %>% dplyr::group_by(Keep.x, Keep.y) %>%  summarise(n = n())
```

```
## # A tibble: 5 x 3
## # Groups:   Keep.x [2]
##   Keep.x Keep.y      n
##    <dbl>  <int>  <int>
## 1      0      0   2935
## 2      0      1    101
## 3      1      0    337
## 4      1      1 165039
## 5      1     NA      1
```

```
join %>%
  dplyr::filter(Keep.x < Keep.y) %>%
  dplyr::select(RateFlag.x, CourseFlag.x, DistFlag.x, RateFlag.y, CourseFlag.y, DistFlag.y) %>%
  dplyr::summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  dplyr::mutate(Source = ifelse(grepl(".x", Flag), "Candidate", "Correct"),
                Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot(aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Observations Kept by Correct")
```



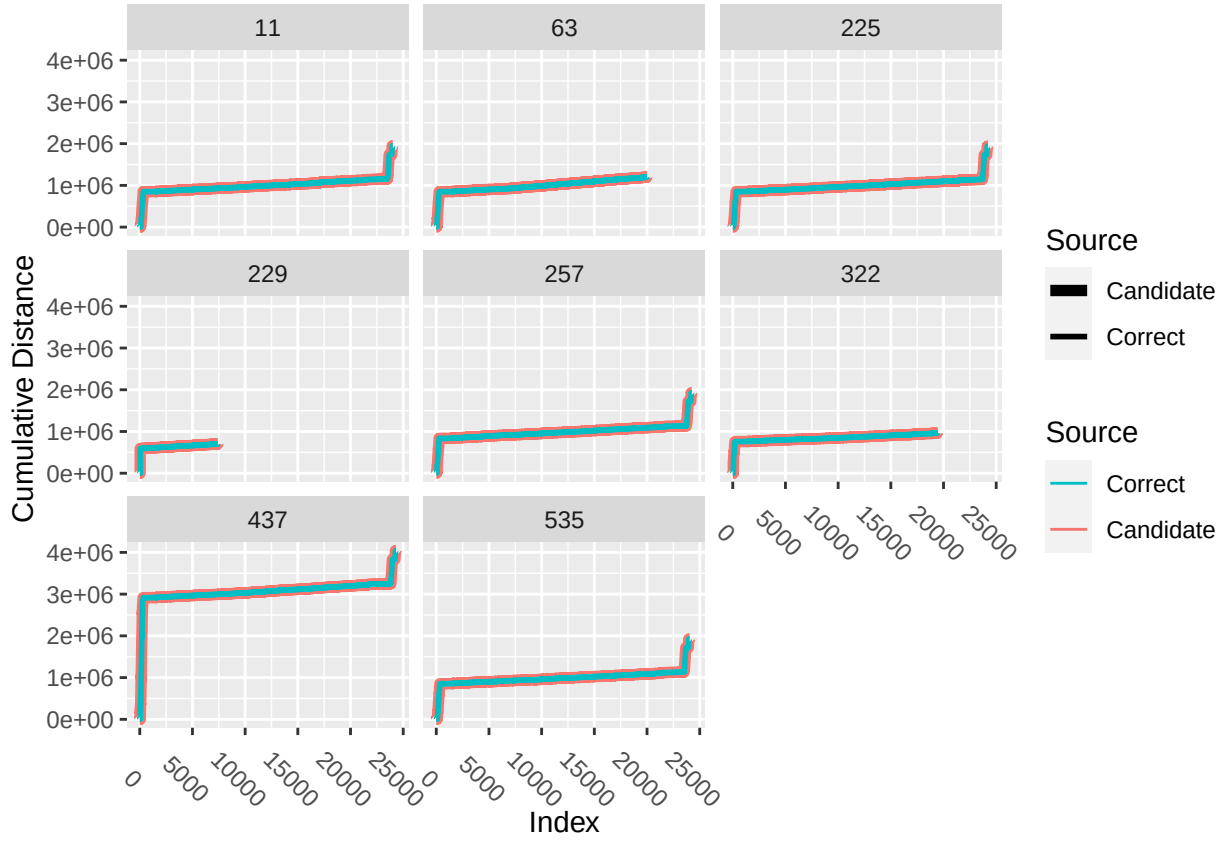

```
join %>%
  dplyr::filter(Keep.y < Keep.x) %>%
  dplyr::select(RateFlag.x, CourseFlag.x, DistFlag.x, RateFlag.y, CourseFlag.y, DistFlag.y) %>%
  dplyr::summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  dplyr::mutate(Source = ifelse(grepl(".x", Flag), "Candidate", "Correct"),
                Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot(aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Observations Kept by Candidate")
```



Cumulative Distance by Cow

```
cumdist <- join %>%
  dplyr::group_by(Cow) %>%
  dplyr::arrange(Index, .by_group=TRUE) %>%
  dplyr::mutate(Distance.y = dplyr::lag(Distance.y,1),
                Distance.x = ifelse(is.na(Distance.x), 0, Distance.x),
                Distance.y = ifelse(is.na(Distance.y), 0, Distance.y),

                cumDist.x = cumsum(Distance.x),
                cumDist.y = cumsum(Distance.y)) %>%
  dplyr::ungroup()

cumdist_candidate <- cumdist %>%
  dplyr::select(Index, Cow, cumDist.x, DistFlag.x) %>%
  dplyr::rename(Flag = DistFlag.x,
                cumDist = cumDist.x) %>%
  dplyr::mutate(Source = "Candidate")

cumdist_correct <- cumdist %>%
  dplyr::select(Index, Cow, cumDist.y, DistFlag.y) %>%
  dplyr::rename(Flag = DistFlag.y,
                cumDist = cumDist.y) %>%
  dplyr::mutate(Source = "Correct")
```

```
plot_data <- dplyr::bind_rows(cumdist_candidate, cumdist_correct)

ggplot(plot_data, aes(x=Index, y=cumDist, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  #geom_point(data=plot_data %>% dplyr::mutate(Flag = ifelse(is.na(Flag), 0, Flag)) %>% dplyr::filter(F
  ylab("Cumulative Distance") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```



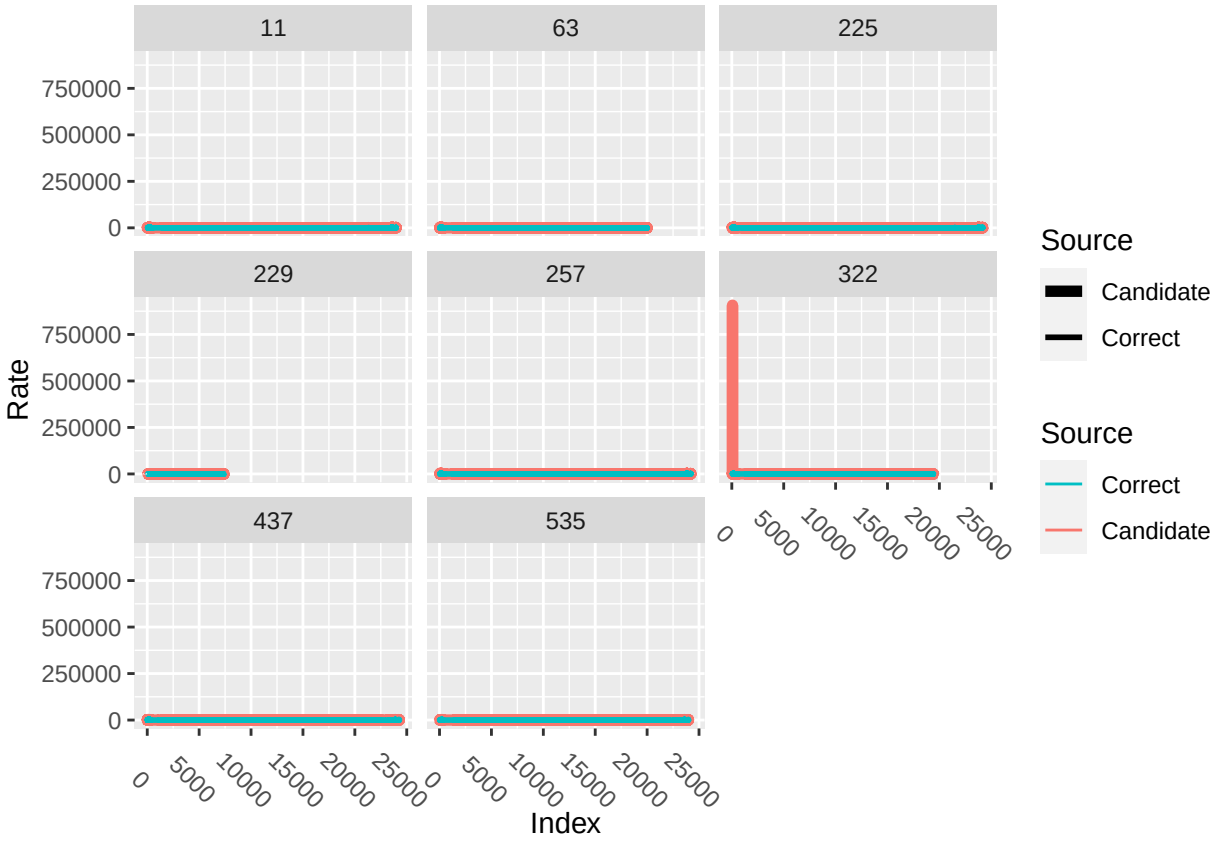
Rate by Cow

```
rate_candidate <- join %>%
  dplyr::select(Index, Cow, Rate.x, RateFlag.x) %>%
  dplyr::rename(Flag = RateFlag.x,
                Rate = Rate.x) %>%
  dplyr::mutate(Source = "Candidate")

rate_correct <- join %>%
  dplyr::select(Index, Cow, Rate.y, RateFlag.y) %>%
  dplyr::mutate(Flag = RateFlag.y,
                Rate = as.numeric(Rate.y)) %>%
  dplyr::mutate(Source = "Correct")
```

```
## Warning: NAs introduced by coercion
```

```
plot_data <- dplyr::bind_rows(rate_candidate, rate_correct)

ggplot(plot_data, aes(x=Index, y=Rate, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  #geom_point(data=plot_data %>% dplyr::mutate(Flag = ifelse(is.na(Flag), 0, Flag)) %>% dplyr::filter(F
  ylab("Rate") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



Course by Cow

```
course_candidate <- join %>%
  dplyr::select(Index, Cow, Course.x, CourseFlag.x) %>%
  dplyr::rename(Flag = CourseFlag.x,
                Course = Course.x) %>%
  dplyr::mutate(Source = "Candidate")

course_correct <- join %>%
  dplyr::select(Index, Cow, Course.y, CourseFlag.y) %>%
  dplyr::rename(Flag = CourseFlag.y,
                Course = Course.y) %>%
  dplyr::mutate(Source = "Correct")
```

```r
plot_data <- dplyr::bind_rows(course_candidate, course_correct)

ggplot(plot_data, aes(x=Index, y=Course, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  #geom_point(data=plot_data %>% dplyr::mutate(Flag = ifelse(is.na(Flag), 0, Flag)) %>% dplyr::filter(F
  ylab("Course") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```